Big Data and Computing Visions

www.bidacv.com

Big. Data. Comp. Vis. Vol. 4, No. 4 (2024) 245–256.

Paper Type: Original Article

Entropic Imprints on Bioinformatics

Ismail A Mageed*

Department of Computer Science, Artificial Intelligence and Electronics, Faculty of Engineering and Digital Technologies, University of Bradford, UK; drismail664@gmail.com.

Citation:

Received: 18 June 2024 *Revised: 14 August 2024* *Accepted: 19 September 2024*	Mageed, I. A. (2024). Entropic imprints on bioinformatics. *Big data and computing visions*, *4*(4), 245-256.

Abstract

The entropic framework is a crucial method in statistical inference that helps scientists create models to describe and predict biological systems, particularly complex networks like gene interactions. Its effectiveness comes from its simple concepts and solid mathematical foundation, making it applicable in various situations. This paper reviews the significant impact of entropy in steering Bioinformatics giant steps ahead through analyzing biological data, such as reconstructing how genes interact and understanding bacterial metabolism and much more. Indeed, the steering hand of entropy in Bioinformatics was arguably in question for many decades until the start of the exploratory era of depicting the phenomenal applicability of entropy in interdisciplinary fields of human knowledge, especially Bioinformatics. Undoubtedly, entropy has a great imprint, which has already changed the way we should think of Bioinformatics based on its multi-faced nature, whether being looked at from an angle of physics, information-theoretic, thermodynamical, chaotic-led approach, and a bird-eye view of a computing perspective. The flow of the current review continues by showcasing the entropic fingerprints on Bioinformatics, resulting in many exceptional discoveries that have enormously added to the existing knowledge. Most importantly, some emerging open problems are provided. Providing these open problems would depict a plethora of issues that need further exploration to build bridges for contemporary Entropic-Bioninformatics Theory. The paper ends with concluding remarks and future research pathways.

Keywords: Information theory, Entropy, Biological data, Modelling, Bioinformatics.

1 | Introduction

This section overviews what we mean by entropy in different settings, combined with a panaoramic spotlight on Bioinformatics.

1.1 | Welcome to Entropy Land

Entropy is a concept that originated in thermodynamics but has since expanded into various fields [1], leading to confusion about its meaning and application. Many different types of entropy have been developed for various purposes [1], which can bewilder researchers due to the lack of clear definitions and relationships

Corresponding Author: drismail664@gmail.com
https://doi.org/10.22105/bdcv.2024.479239.1201

between them. *Fig. 1* [1] presents a timeline that organizes different types of entropies. This timeline uses a logarithmic scale [1], which helps to display a wide range of values more clearly. The timeline of the universe of entropies refers [1] to a visual representation that organizes different types of entropy concepts and their development over time. This timeline helps to illustrate how various entropy measures have emerged [1], their mathematical definitions, and how they relate to one another.

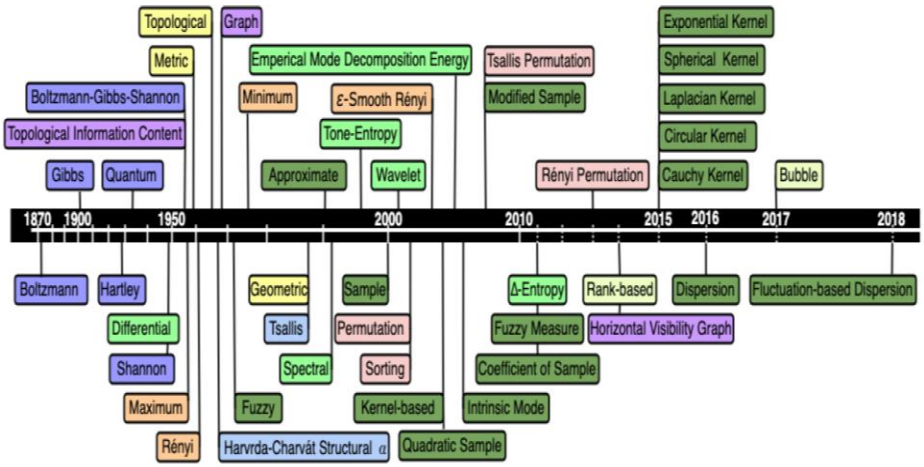

Fig. 1. Timeline illustrating the evolution and application of entropy across fields.

Fig. 2 [1] likely shows how various entropy measures, such as sample entropy and multiscale entropy, relate to each other and their roles in analyzing complex systems. Understanding this framework helps researchers grasp the broader implications of entropy in different scientific contexts. In this context [1], it includes various types of entropy, such as conditional entropy, which looks at how knowing one variable can reduce uncertainty about another. Additionally, it highlights different methods for quantifying the complexity and predictability of time-series data, like approximate entropy and sample entropy, which are used across various fields to analyze patterns and behaviours in data.

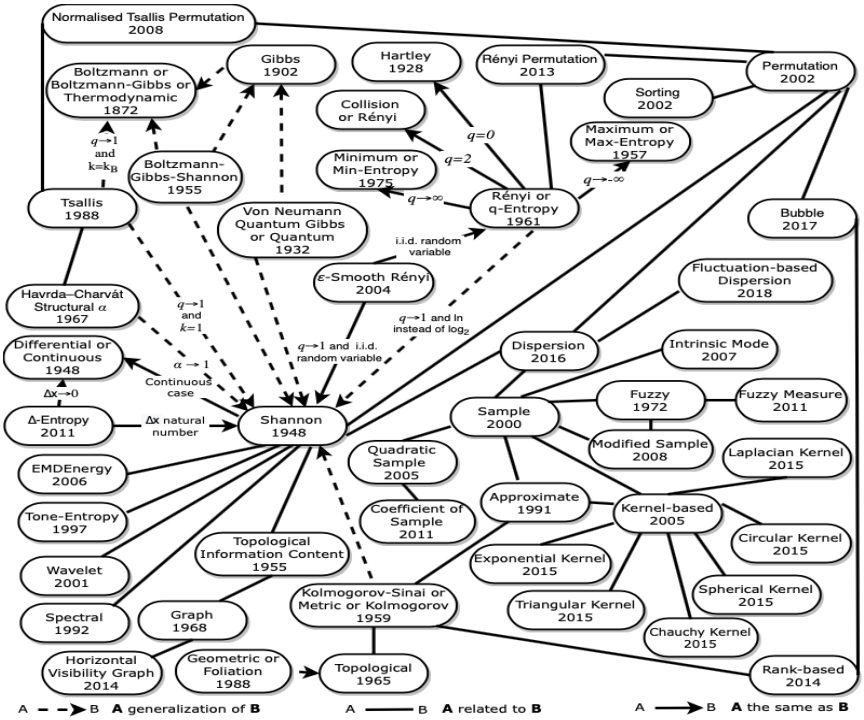

Fig. 2. The entropy universe refers to the broad concept of entropy in information theory, which measures uncertainty or randomness in data.

The Shannonian entropy functional [2], denoted as $H(p)$, is a mathematical formula that measures the amount of information in a system based on the probabilities of that system's different states (or configurations). It is

particularly useful for understanding systems with short-range interactions, where the behaviour of one part of the system is influenced mainly by its immediate neighbours.

$H(p)$ reads as:

$$H(p) = -\sum_{n=0}^{\infty} p(n)\ln p(n). \quad (1)$$

Such that $p(n)$ define the short-range state probabilities of the underlying system.

When considering systems with "long-range" interactions, a different functional [2], namely Ismail's entropy, $H_{(q,UG)}$ is proposed to account for the more complex relationships between states, allowing for a broader analysis of how these interactions affect the overall system. $H_{(q,UG)}$ is the ultimate generalization to $H(p)$ *Eq. (1)*, where

$$H_{(q,UG)} = \sum_{n=0}^{\infty} \varphi\left((p(n))^{q}, a_1, a_2, .., a_k\right), k \leq n. \quad (2)$$

φ serves as any uniquely- defined function and $a_1, a_2, .., a_k$ are any universal parameters with $1 > q > 0.5$.

It is worth noting that [2–8], $H_{(q,UG)}$, ultimately generalizes most entropies in the literature, such as Renyi's and Tsallis' entropies, etc.

1.2 | Bioinformatics

Many people [9] today think bioinformatics is a new field created to help analyze data from next-generation sequencing. However, bioinformatics actually began over 50 years ago, focusing initially on protein sequence analysis before expanding to DNA analysis as technology improved. As sequencing technology advanced and data volumes grew [9], [10], bioinformatics became essential in biology, leading to new subfields and prompting universities to incorporate it into biology education.

Automated Edman peptide sequencing, as depicted by *Fig. 3* [9], is a method used to determine the sequence of amino acids in a protein. In this process [9], the first amino acid at the beginning of the peptide chain (the N-terminus) is labelled with a chemical called phenylisothiocyanate (PITC) and then removed by changing the pH. This process is repeated to identify each amino acid one at a time [9], allowing scientists to piece together the entire protein sequence more efficiently.

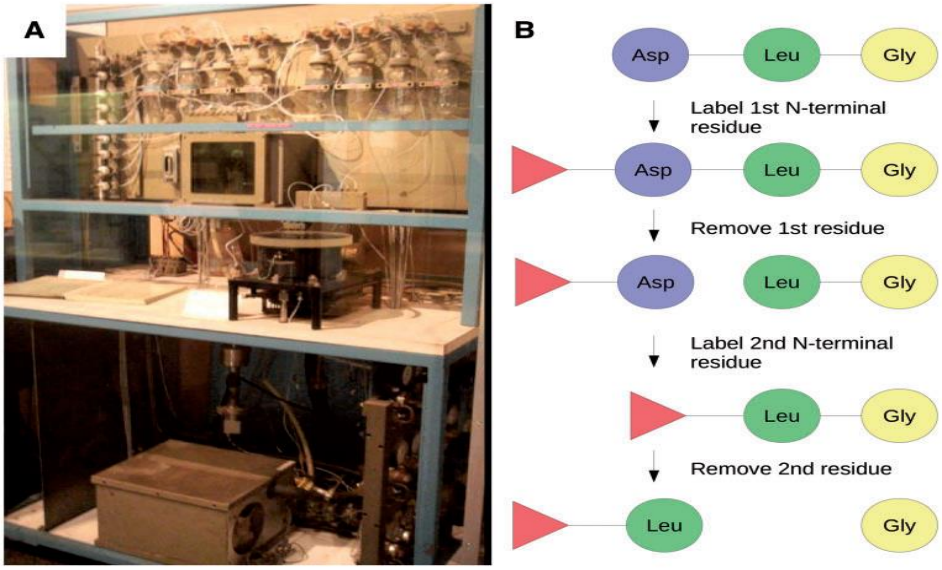

Fig. 3. Automated Edman peptide sequencing.

The Needleman-Wunsch global alignment algorithm [9], [10] as showcased by *Fig.4* [9], is a method used in bioinformatics to find the best alignment between two DNA or protein sequences. It works by creating a scoring matrix that assigns points for matches, penalties for mismatches, and additional penalties for gaps in the sequences. The algorithm calculates scores for each cell in the matrix based on the best scores from neighbouring cells, ultimately leading to the optimal alignment of the sequences, as shown in the example with ATCG and ATG sequences.

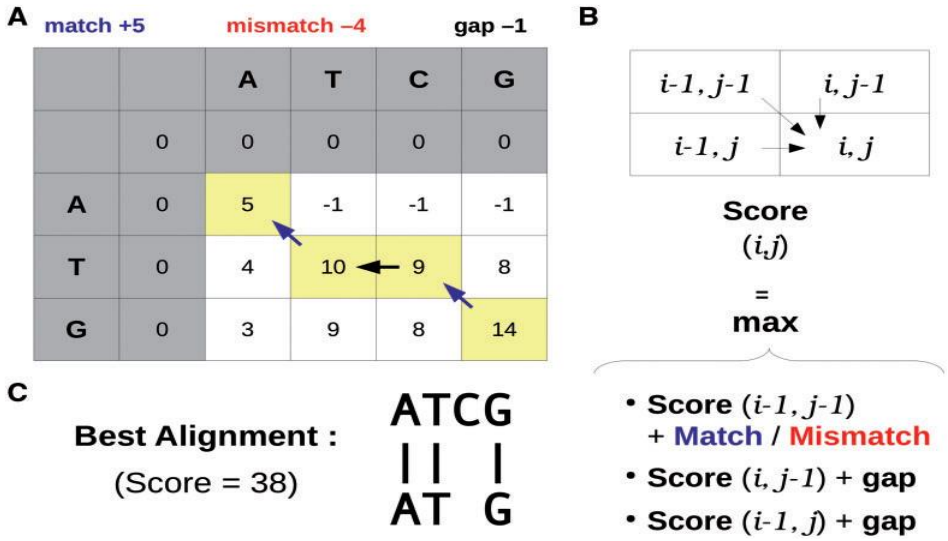

Fig. 4. The needleman-wunsch global alignment algorithm.

Hierarchical shotgun sequencing and Whole Genome Shotgun (WGS) sequencing are two different methods used to sequence the human genome [9], as illustrated by *Fig.5*, representing a competition between public (NIH) and private (Celera) projects. The NIH team thought that WGS was too complicated for the large size of the human genome, while Celera's team, led by Craig Venter, believed it was possible and could simplify the process compared to hierarchical shotgun sequencing. They argued that with the right computer algorithms and enough computing power, WGS could effectively handle the challenges of sequencing such a vast amount of genetic information.

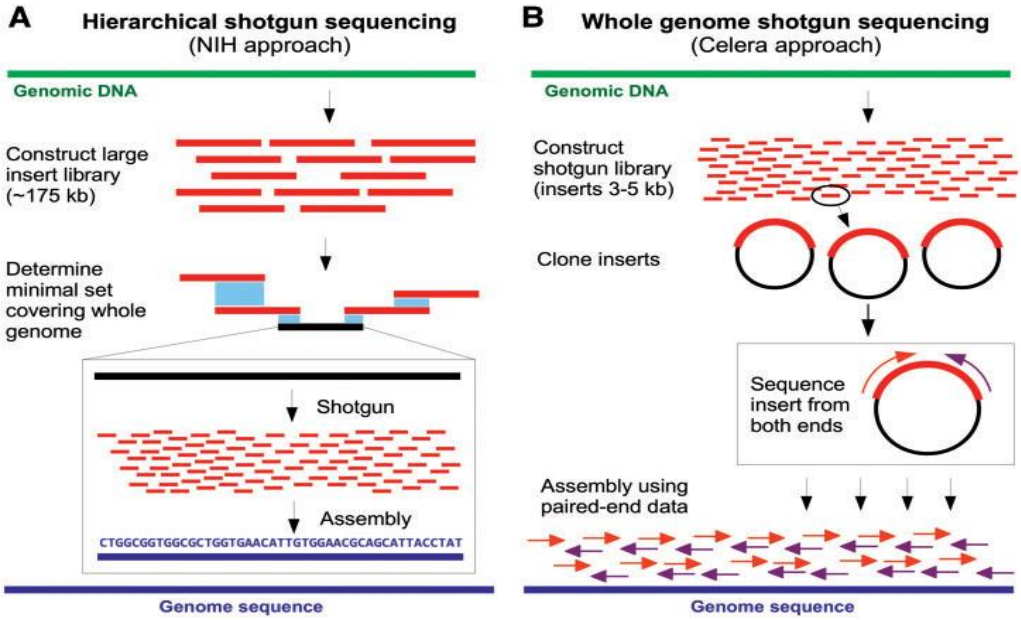

Fig. 5. Hierarchical shotgun sequencing and WGS sequencing.

The historical timeline [11] of circRNA research highlights key discoveries and advancements in understanding circular RNAs (circRNAs) as visualized in *Fig. 6* [11]. It shows how circRNAs were first identified in plant viruses and later found in eukaryotic cells and humans, with significant milestones including

their ability to synthesize proteins and act as miRNA sponges. The timeline also emphasizes the development of experimental and bioinformatics tools that have accelerated circRNA research [11], revealing their potential roles in biology and as cancer biomarkers.

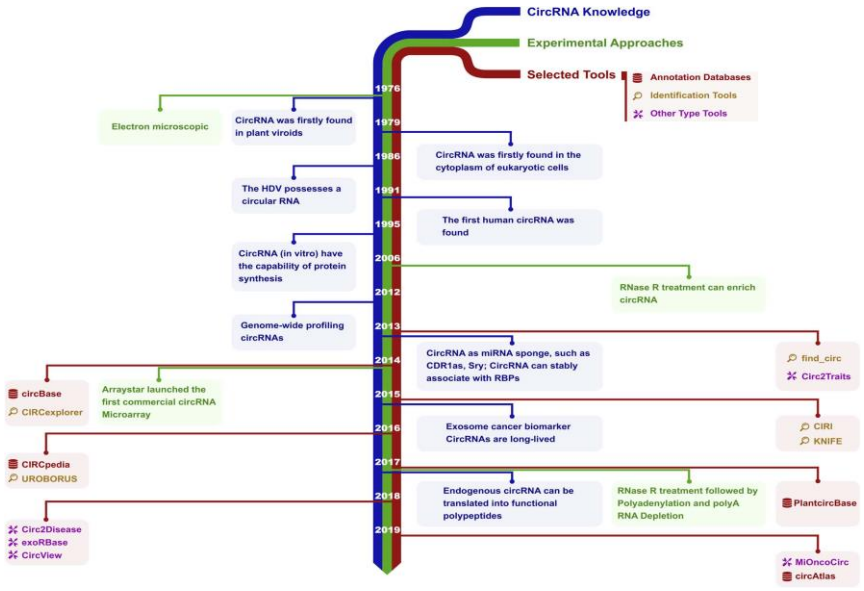

Fig. 6. key discoveries and advancements in understanding circular RNAs (circRNAs).

The flowchart for protein three-dimensional structure prediction outlines the process of determining the 3D shape of a protein based on its amino acid sequence, see *Fig.7* [11]. It starts with the target protein sequence and can use different methods depending on available templates or constraints, with homology-based models being the most accurate and ab-initio models being less reliable. The initial model is built through sequence alignment [11], and it may require further refinement to fill in any missing parts and improve its quality before arriving at the final predicted structure.

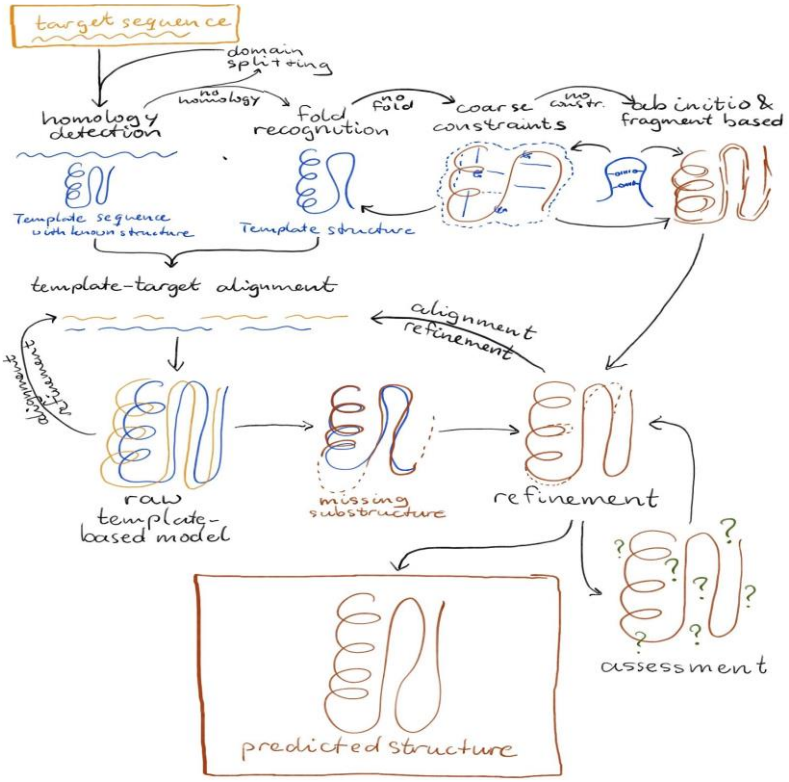

Fig. 7. The process of determining the 3D shape of a protein based on its amino acid sequence.

A fragment library [12] is created by taking short pieces (3 and 9-residue fragments) from known protein structures in the Protein Data Bank (PDB) and grouping them based on their similar shapes. When predicting

the structure of a new protein [12], the target sequence is also broken down into overlapping fragments, which are then matched with the fragments in the library to find suitable structural pieces. This process helps build a model of the protein's structure by combining these fragments [12], which is further refined to find the best possible representation of the protein.

2 | Entropic Impact on Bioinformatics

Entropy [13], in the context of thermodynamics, is often described as a measure of disorder or randomness in a system. It quantifies [13] the number of different ways that the components of a system can be arranged while still achieving the same overall state, such as the distribution of cells in different states.

The maximum entropy principle suggests that when inferring a probability distribution under certain constraints [13], the best choice is the one that maximizes entropy, reflecting the most uncertainty or randomness consistent with the given information. The Maximum Entropy (MaxEnt) principle [13] is a statistical method used to analyze biological data by making the least biased predictions based on available information. In the context of gene interaction networks, it helps infer how genes interact with each other using expression data, while for metabolic flux patterns [13], it analyzes growth rate distributions in bacteria to understand how nutrients are processed. Each application involves taking empirical data as input, formulating a problem using the maximum entropy approach, and then deriving meaningful biological insights from the results, as shown by *Fig.8* [13].

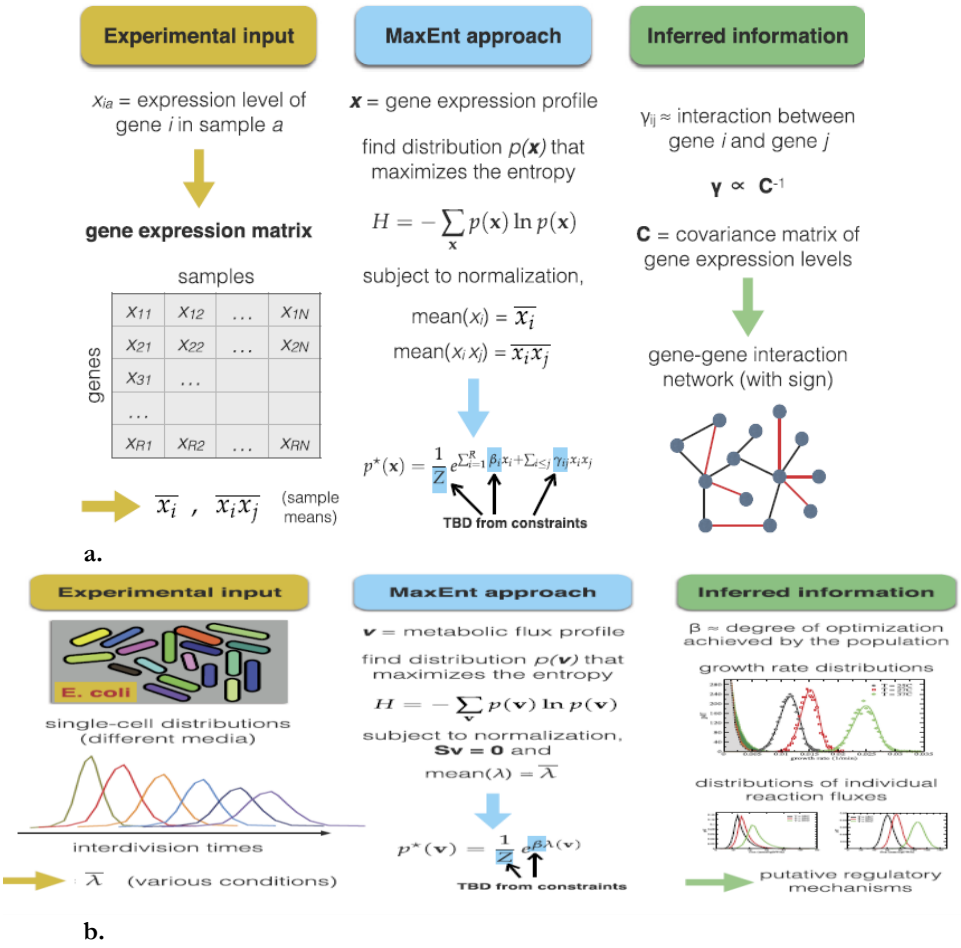

Fig. 8. Applications of maximum entropy principle in biological data analysis: a. Gene interaction networks from expression data; b. Inferring metabolic flux patterns from bacterial growth rate distributions.

A thorough discussion of various methods and tools used in bioinformatics to measure the complexity of DNA sequences was undertaken by [14], focusing on aspects like information and entropy. These complexity

measures help identify important regions in the genome, such as low-complexity regions, which play crucial roles in chromosome function and gene regulation. Compression techniques and machine learning to analyze genetic sequences were also highlighted [14], including applications for both DNA and protein sequences, emphasizing the significance of low-complexity regions in evolution and biological functions.

Fig.9 [14] illustrates different approaches, like analyzing the frequency of nucleotides and using entropy measures, as well as tools like CorGen and NeSSie that help identify patterns in DNA. Additionally, it highlights how certain regions of DNA, such as those involved in gene regulation, tend to have lower complexity compared to protein-coding regions, which can provide insights into how genes are expressed and regulated.

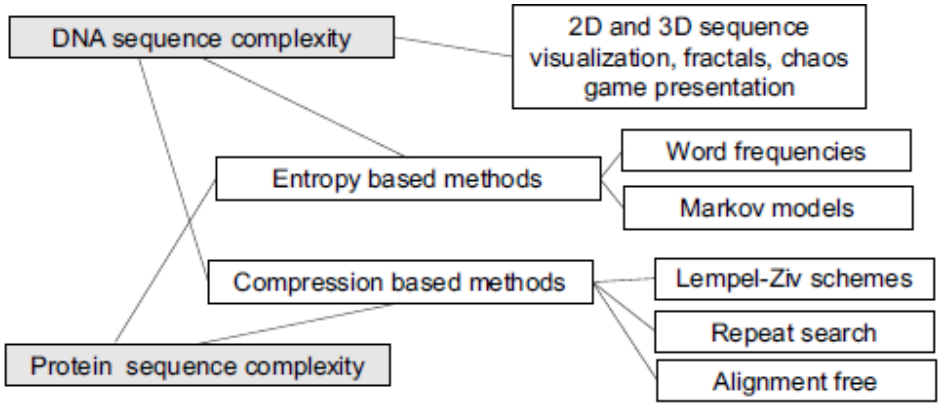

Fig. 9. Visualization of various methods for estimating the complexity of DNA sequences, which can help researchers understand genetic information better.

The complexity estimates for different types of genome sequences, such as exons (coding regions), introns (non-coding regions) [14], and regulatory sequences, support the idea that more genetic information in a sequence leads to higher complexity. Research shows [14] that exons generally have higher complexity than introns because introns often contain simpler, repetitive sequences. This understanding is important for analyzing genomes, as it helps scientists identify patterns and functions within the genetic material.

GxG interactions [15] refer to how one genetic variant (like a single nucleotide polymorphism or SNP) can influence the effect of another genetic variant on a specific trait, such as the presence or absence of a disease. The study primarily focuses on binary traits, coded as 0 for controls and 1 for cases, but also considers quantitative traits. Different methods exist to detect these interactions, often using data from either family members or independent individuals, and the article aims to provide a general framework for understanding these interactions in genetics. The conditional entropy *Fig. 10* [15] of a random variable given another variable quantifies the uncertainty of a random variable when the other variable is known, but the joint entropy *Fig. 11* [15] offers the uncertainty of two random variables simultaneously. Ultimately, the mutual information *Fig. 12* [15] between two variables shows how knowing the other one reduces the uncertainty of one.

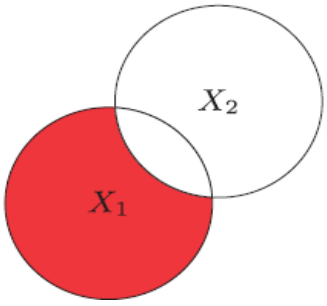

Fig. 10. The mutual information can be conditioned on a third variable, just like entropy, to produce Conditional Mutual Information (CMI).

In *Fig.10.* Conditional entropy of the variable X_1 given the variable X_2 measures the amount of uncertainty remaining about X_1 when we know the value of X_2. It is calculated using the probabilities of X_1 and X_2

together, compared to the probability of just X_2 . Essentially, it helps us understand how much information about X_1 can be gained by knowing X_2 , and is represented mathematically as the difference between the joint entropy of X_1 and X_2 and the entropy of X2 alone.

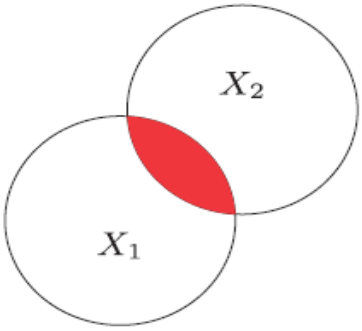

Fig. 11. Mutual information between two variables, X_1 and X_2 , measures how much knowing the value of one variable tells us about the other.

It quantifies the amount of shared information between the two variables, indicating how much uncertainty is reduced about one variable when the other is known. In simpler terms, it helps us understand the relationship and dependence between X_1 and X_2 in a statistical context.

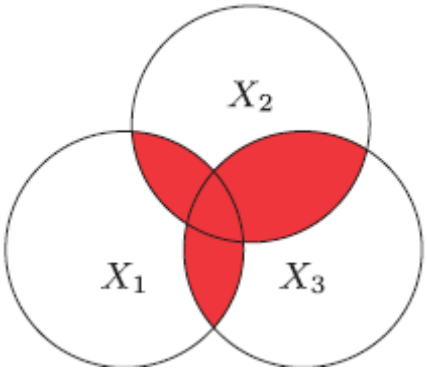

Fig. 12. Total Correlation Information (TCI) of three random variables measures the overall dependence among those variables.

It quantifies how much information is shared among all three variables compared to what each variable contributes individually. Essentially, TCI helps us understand the combined relationships and interactions between the variables more comprehensively.

Reducing the genes [16] in gene expression datasets that are superfluous or unnecessary can significantly lower the cost of cancer classification. In order to enhance the classification performance of gene expression data [16], a feature selection approach based on neighbourhood entropy-based uncertainty measures is presented in this study. First [16], the effectiveness of neighbourhood entropy-based uncertainty measures in quantifying uncertainty and removing noise from gene expression datasets is examined. Next [16], in order to completely characterize the decision-making capacity of characteristics in neighbourhood decision systems, the neighbourhood credibility degree and the neighbourhood coverage degree are added to decision neighbourhood entropy and mutual information, which have been shown to be nonmonotonic.

Depending on the number of genes in the chosen subset and the precision of the categorization in every high-dimensional gene expression dataset. A heuristic reduction technique [16] for cancer classification is developed to effectively reduce the computational complexity and enhance the classification performance of gene expression data by employing the Fisher score algorithm prior to lowering the dimensionality of gene expression datasets. The experimental findings demonstrate that, in gene expression datasets [16], the suggested algorithm is able to identify a small, useful subset of genes and achieve high classification accuracy. Nevertheless [16], not all high-dimensional gene expression datasets can be classified with an optimal balance between the size of the selected gene subset and classification accuracy using our suggested strategy.

Information can be understood as reducing uncertainty [17–31] or entropy, in a receiver when they receive a message. In biology, for example, when mRNA reaches a ribosome, it provides specific instructions for assembling proteins, reducing the randomness of how amino acids are put together. The difference in entropy before and after receiving the mRNA indicates how much information has been transmitted, with maximum information occurring when no uncertainty is left after receiving the message.

Relative entropy [17] at each position in a biological sequence measures the amount of information present at that specific spot. This concept is used to create sequence logos, visually representing biological sequences by showing the relative importance of symbols (like nucleotides or amino acids) at each position. In a sequence logo [17], the height of each letter indicates how much information that position contributes, allowing researchers to see which parts of the sequence are more significant easily, see *Fig.13* [17].

The phrase entropy in alignment reveals information at position according to the positional paradigm refers to how the variability (or entropy) in DNA or protein sequences can indicate the importance of specific positions in those sequences [17]. In the context of the tumour protein p53 [17], two 'sequence logos' are created to represent the information from different parts of the protein visually: the DNA Binding Domain (DBD) and the intrinsically disordered N-Terminal Domain (NTD). These logos help researchers understand which positions in the sequence are more conserved and likely to be functionally important, with the x-axis showing specific positions in the human p53 sequence. As seen in *Fig.13* B, a multiple sequence alignment is a technique used to compare and align sequences from different species to identify similarities and differences. In the case of the p53 protein, the alignment shows the folded DNA binding domain at the top and the intrinsically disordered N-terminal domain at the bottom. The key difference is that the folded domain has high positional conservation, meaning its sequence is more stable across species, while the intrinsically disordered region has low conservation, indicating it varies more and is less structured.

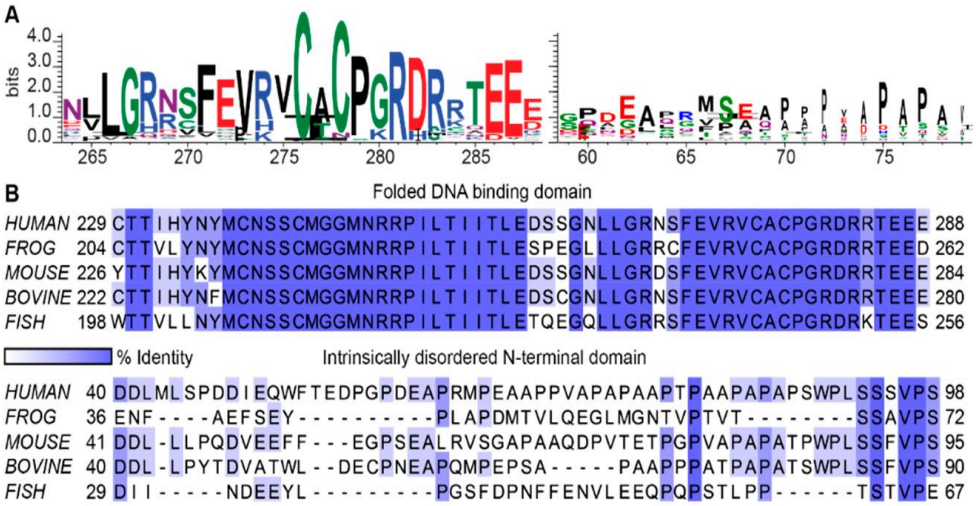

Fig. 13. The positional paradigm states that information at position is revealed via entropy in alignment.

3 | Open Problems

I. Following [13], there are risen de nove challenges, which include understanding the data and using it to create predictive models or new design methods. Essentially, the complexity of biological datasets makes it tough for researchers to analyze and apply their findings effectively.

II. The undertaken line of enquiry [15] suggests that researchers need to develop a clear and consistent way to define interactions in genetics using the concept of entropy, which measures uncertainty or information. This involves figuring out which existing definitions are similar, repetitive, or conflicting. Some initial work has already been done to compare different measures of interaction to see how they relate to interactions found using a logistic regression model.

III. The systematic literature search [15] revealed that it is often unclear how to create accurate statistical tests using the proposed methods. Simply substituting probabilities with frequencies can lead to biased results or estimators that take a long time to provide reliable answers. Therefore [14], researchers should aim to develop estimators that are consistent (reliable over time) and ideally unbiased (not skewed in any direction).

IV. The systematic literature search [15] uncovered that researchers often do not know the statistical distributions of test results when comparing two hypotheses: the null hypothesis (which suggests no effect) and the alternative hypothesis (which suggests an effect). Because of this uncertainty, there is a need for estimators that can analyze the long-term behaviour of these test statistics as sample sizes increase, known as their "asymptotic behaviours." Understanding these behaviours is important for making accurate conclusions in genetic studies.

V. The work of [15] discussed a method for analyzing genetic interactions using a measure called IDIgnac, which helps determine whether two genetic variants work together (synergy) or do not affect each other (redundancy). This reflects that different statistical tests are used to explore these interactions and highlights the need for better definitions and methods to assess genetic relationships accurately. Finally, it points out that many existing methods are inconsistent or unclear, indicating that more research is needed to improve these statistical approaches in genetic studies.

VI. The authors [15] emphasized the need for careful consideration when using information theory-based estimators to study gene-gene (GxG) interactions. They recommend clarifying the types of interactions defined by the chosen methods, as simpler approaches may overlook important signals. Additionally, they suggest starting with second-order interactions before moving to more complex ones, as this can help manage the challenges of analyzing high-dimensional data effectively.

VII. Based on [16], new search strategies and more effective uncertainty measures based on neighbourhood rough sets should be investigated to improve further the classification performance and computational efficiency of the proposed algorithms for cancer classification and to make our algorithms more appropriate for application areas, such as big data mining, pattern recognition, and bioinformatics for biomarker discovery.

VIII. Building on [16], new search tactics and more potent uncertainty measures based on neighbourhood rough sets ought to be looked into to enhance the computational efficiency and classification performance of the suggested algorithms for cancer classification and to suit better our algorithms for use in domains like big data mining, pattern recognition, and bioinformatics for biomarker discovery.

Acknowledgments

Acknowledgements enable you to thank all those who have helped carry out the research. Careful thought needs to be given concerning those whose help should be acknowledged and in what order. The general advice is to express your appreciation concisely and to avoid strong emotive language.

Author Contributaion

Ismail A. Mageed conceptualized the study, conducted the literature review, analyzed the implications of entropy in bioinformatics, and drafted the manuscript..

Funding

This work is not funded by any organization.

Data Availability

No new data were generated.

Conflicts of Interest

The author declares no conflict of interest.

References

[1] Ribeiro, M., Henriques, T., Castro, L., Souto, A., Antunes, L., Costa-Santos, C., & Teixeira, A. (2021). The entropy universe. *Entropy*, *23*(2). DOI:10.3390/e23020222

[2] Mageed, I. A., & Zhang, Q. (2022). An information theoretic unified global theory for a stable M/G/1 queue with potential maximum entropy applications to energy works. *2022 global energy conference (gec)* (pp. 300–305). IEEE. https://ieeexplore.ieee.org/abstract/document/9986719/

[3] Mageed, I. A., & Zhang, Q. (2023). Threshold theorems for the tsallisian and rényian (TR) cumulative distribution functions (CDFs) of the heavy-tailed stable M/G/1 queue with tsallisian and rényian entropic applications to satellite images (SIs). *Electronic journal of computer science and information technology*, *9*(1), 41–47. https://ejcsit.uniten.edu.my/index.php/ejcsit/article/view/161

[4] Mageed, I. A., & Zhang, Q. (2023). Formalism of the rényian maximum entropy (RMF) of the stable M/G/1 queue with geometric mean (GeoM) and shifted geometric mean (SGeoM) constraints with potential geom applications to wireless sensor networks (WSNs). *Electronic journal of computer science and information technology*, *9*(1), 31–40.

[5] Mageed, I. A., & Zhang, Q. (2022). An introductory survey of entropy applications to information theory, queuing theory, engineering, computer science, and statistical mechanics. *2022 27th international conference on automation and computing (ICAC)* (pp. 1–6). IEEE. https://ieeexplore.ieee.org/abstract/document/9911077/

[6] Mageed, I. A., & FRSS, I. (2023). *Where the mighty trio meet: information theory (IT), pathway model theory (PMT) and queueing theory (QT)* [presentation]. 39th annual uk performance engineering workshop (p. 8). https://www.researchgate.net/profile/Nigel-Thomas-8/publication/376305981_UKPEW_2023/links/6571edc2cbd2c535ea04874e/UKPEW-2023.pdf#page=10

[7] Mageed, I. A. (2023). Cosistency axioms of choice for Ismail's entropy formalism (IEF) Combined with information-theoretic (IT) applications to advance 6G networks. *European journal of technique (ejt)*, *13*(2), 207–213. https://dergipark.org.tr/en/pub/ejt/issue/80301/1299311

[8] Mageed, I. A. (2023). Fractal dimension (Df) of Ismail's fourth entropy ($HIV^{(q,a}1, a2,..,ak)$) with fractal applications to algorithms, haptics, and transportation. *2023 international conference on computer and applications (ICCA)* (pp. 1–6). IEEE. https://ieeexplore.ieee.org/abstract/document/10401780/

[9] Gauthier, J., Vincent, A. T., Charette, S. J., & Derome, N. (2018). A brief history of bioinformatics. *Briefings in bioinformatics*, *20*(6), 1981–1996. DOI:10.1093/bib/bby063

[10] Lesk, A. (2019). *Introduction to bioinformatics*. Oxford university press.

[11] Chen, L., Wang, C., Sun, H., Wang, J., Liang, Y., Wang, Y., & Wong, G. (2020). The bioinformatics toolbox for circRNA discovery and analysis. *Briefings in bioinformatics*, *22*(2), 1706–1728. DOI:10.1093/bib/bbaa001

[12] Patel, B., Singh, V., & Patel, D. (2019). Structural Bioinformatics. In *Essentials of bioinformatics, volume i: understanding bioinformatics: genes to proteins* (pp. 169–199). Cham: Springer International Publishing. DOI: 10.1007/978-3-030-02634-9_9

[13] De Martino, A., & De Martino, D. (2018). An introduction to the maximum entropy approach and its application to inference problems in biology. *Heliyon*, *4*(4).

[14] Orlov, Y. L., & Orlova, N. G. (2023). Bioinformatics tools for the sequence complexity estimates. *Biophysical reviews*, *15*(5), 1367–1378. DOI:10.1007/s12551-023-01140-y

[15] Ferrario, P. G., & König, I. R. (2016). Transferring entropy to the realm of GxG interactions. *Briefings in bioinformatics*, *19*(1), 136–147. DOI:10.1093/bib/bbw086

[16] Sun, L., Zhang, X., Qian, Y., Xu, J., & Zhang, S. (2019). Feature selection using neighborhood entropy-based uncertainty measures for gene expression data classification. *Information sciences*, *502*, 18–41.

[17] Pritišanac, I., Vernon, R. M., Moses, A. M., & Forman Kay, J. D. (2019). Entropy and information within intrinsically disordered protein regions. *Entropy*, *21*(7), 662. https://www.mdpi.com/1099-4300/21/7/662

[18] Zenil, H. (2020). A review of methods for estimating algorithmic complexity: options, challenges, and new directions. *Entropy*, *22*(6). DOI:10.3390/e22060612

[19] Roscoe, S., Khatri, M., Voshall, A., Batra, S., Kaur, S., & Deogun, J. (2022). Formal concept analysis applications in bioinformatics. *ACM computing surveys*, *55*(8), 1–40. https://dl.acm.org/doi/abs/10.1145/3554728

[20] Xie, Y., Peng, X., & Li, P. (2024). MIWE: detecting the critical states of complex biological systems by the mutual information weighted entropy. *BMC bioinformatics*, *25*(1), 44. https://link.springer.com/article/10.1186/s12859-024-05667-z

[21] Löchel, H. F., & Heider, D. (2021). Chaos game representation and its applications in bioinformatics. *Computational and structural biotechnology journal*, *19*, 6263–6271. https://www.sciencedirect.com/science/article/pii/S2001037021004736

[22] Guo, Z., Liu, J., Wang, Y., Chen, M., Wang, D., Xu, D., & Cheng, J. (2024). Diffusion models in bioinformatics and computational biology. *Nature reviews bioengineering*, *2*(2), 136–154.

[23] Lyu, C., Chen, L., & Liu, X. (2024). Detecting tipping points of complex diseases by network information entropy. *Briefings in bioinformatics*, *25*(4). https://academic.oup.com/bib/article-abstract/25/4/bbae311/7705534

[24] Bonnici, V., & Manca, V. (2015). Infogenomics tools: a computational suite for informational analysis of genomes. *Journal of bioinformatics and proteomics review*, *1*(1), 7–14. https://air.unipr.it/handle/11381/2901603

[25] Bonidia, R. P., Sampaio, L. D. H., Domingues, D. S., Paschoal, A. R., Lopes, F. M., de Carvalho, A. C. P. L. F., & Sanches, D. S. (2021). Feature extraction approaches for biological sequences: a comparative study of mathematical features. *Briefings in bioinformatics*, *22*(5), bbab011. DOI:10.1093/bib/bbab011

[26] Shi, Y., Wang, M., Shi, W., Lee, J.-H., Kang, H., & Jiang, H. (2019). Accurate and efficient estimation of small P-values with the cross-entropy method: applications in genomic data analysis. *Bioinformatics*, *35*(14), 2441–2448. https://academic.oup.com/bioinformatics/article-abstract/35/14/2441/5232222

[27] Li, Y., Huang, C., Ding, L., Li, Z., Pan, Y., & Gao, X. (2019). Deep learning in bioinformatics: introduction, application, and perspective in the big data era. *Methods*, *166*, 4–21. https://www.sciencedirect.com/science/article/pii/S1046202318303256

[28] Alansari, Z., Anuar, N. B., Kamsin, A., Soomro, S., & Belgaum, M. R. (2019). Evaluation of iot-based computational intelligence tools for dna sequence analysis in bioinformatics. *Progress in advanced computing and intelligent engineering* (pp. 339–350). Singapore: Springer Singapore. https://doi.org/10.1007/978-981-13-0224-4_31

[29] Dan, S., Kar, M. B., Majumder, S., Roy, B., Kar, S., & Pamucar, D. (2019). Intuitionistic type-2 fuzzy set and its properties. *Symmetry*, *11*(6). https://www.mdpi.com/2073-8994/11/6/808

[30] Mandal, A. K., Sarma, P. K. D., & Dehuri, S. (2023). A study of bio-inspired computing in bioinformatics: a state-of-the-art literature survey. *The open bioinformatics journal*, *16*(1). https://openbioinformaticsjournal.com/VOLUME/16/ELOCATOR/e187503622305100/FULLTEXT/

[31] Deepthi, P. N., Anitha, R., & Swathi, K. (2019). Retracted: a review on bioinformatics using data mining techniques. *Journal of physics: conference series* (Vol. 1228, p. 12023). IOP Publishing.